Chapter 8

Multiple Regression Models in Forecasting

Multiple regression analysis extends the power of simple regression to deal with more complex business situations where sales depend on two or more causal influences. Computers handle the calculations readily. Therefore, this chapter concentrates on (1) framing multiple regression hypotheses to test and (2) analyzing the meaning of computer generated results.

8.1 Functional Form of Multiple Linear Regression

To expand the simple linear regression model to include two or more explanatory independent variables (causal influences), the estimated regression equation is extended from:

$$Y_i = b_1 + b_2X_i + \epsilon_i \tag{8.1}$$

to:

$$Y_i = b_1 + b_2X_{i2} + b_3X_{i3} + \ldots + b_jX_{ij} + \ldots + b_kX_{ik} + \epsilon_i \tag{8.2}$$

In an elementary sense, b_1 is the predicted value of sales, Y_{ic}, where all the X_{ij}'s are zero. Since this case is not likely to happen in sales forecasting, we regard b_1 more generally as a height adjustment constant. The other coefficients (b_2, $b_3, \ldots, b_k$) measure the effect each independent variable has on sales. More rigorously, b_2 measures the changes in sales, Y_{ic}, per unit change in X_{i2}, with all other X_{ij}'s held constant (in the statistical sense). Of course ϵ_i continues to represent all available information on the ways in which the fitted regression model fails to properly explain the observed dispersion in the sales variable, i.e., $\epsilon_i = (Y_i - Y_{ic})$.

As in simple regression, for multiple analysis we first determine the economic and business factors which cause changes in sales. Next, we find the appropriate numerical time series which embody these factors. Finally, we describe mathematically the structure by which the explanatory time-series variables are expected to affect sales. We might begin with the simple linear form of Equation 8.1 but may wish to extend it to the multiple analysis embodied in Equation 8.2.

Similarly, to complete the specification of the multiple regression model, we extend the underlying assumptions for analysis, as follows:

1. *Linearity*. The multiple regression equation is based on the premise of a linear relationship between sales, Y_i, and all the explanatory variables, X_{ij}'s.
2. *Normality*. For each set of the explanatory $X_{.j}$ values, the dispersion of points about the regression line is normally distributed–i.e., the ϵ_i, are normally dispersed.
3. *Homoscedasticity*. There is a uniform scatter of points around the regression line–i.e., constant variance throughout the range of $X_{.j}$ observations.
4. *Independence*. The values of the ϵ_i (residual error) are statistically independent (i.e., independence of successive observations) of one another such that the expected value of ϵ_i is zero and its variance is constant for all i observations.
5. *Non-multicollinearity*. No linear relationships exist between (among) any of the explanatory, $X_{.j}$, variables.

8.2 Case Study: Process Control Company

Table 8.1 presents quarterly seasonally adjusted sales for Process Control Company in addition to time and two explanatory variables from the U.S. economy: (1) New plant and equipment expenditures for manufacturing durable goods industries and (2) Corporate profits and inventory valuation. These explanatory variables were selected since Process Control's primary markets are inside the U.S. economic structure; thus consumption of the company's output is adjudged a function of these causal influences.

For three explanatory variables, the general form of the forecasting equation is:

$$Y_c = b_1 + b_2X_2 + b_3X_3 + b_4X_4 \tag{8.3}$$

Figure 8.1 shows the output of a regression calculation on a time-sharing computer console. The following narrative describes the computer output and is indexed according to

TABLE 8.1 -- Process Control Company: Seasonally Adjusted Data for Multiple Regression Analysis

(1) Quarter Year	(2) Y Sales (1/10 mil dollars)	(3) Explanatory Variables from United States Economy: Quarter Year	(4) X_2 New Plant& Equip Expend Mfg., Dur. (bil. dol.)	(5) Quarter Year	(6) X_3 Corp. Profits & Inven. Value (bil.dol.)	(7) X_4 Time Trend
3-1965	226	1-1966	13.28	1-1965	73.1	1
4-	245	2-	13.98	2-	74.4	2
1-1966	254	3-	14.18	3-	76.5	3
2-	285	4-	14.58	4-	80.3	4
3-	261	1-1967	14.46	1-1966	81.5	5
4-	249	2-	14.26	2-	82.1	6
1-1967	242	3-	13.92	3-	82.5	7
2-	225	4-	13.71	4-	83.7	8
3-	235	1-1968	14.11	1-1967	78.3	9
4-	225	2-	13.51	2-	78.0	10
1-1968	216	3-	14.47	3-	78.4	11
2-	224	4-	14.39	4-	80.0	12
3-	245	1-1969	15.47	1-1968	81.1	13
4-	300	2-	15.98	2-	85.4	14
1-1969	327	3-	16.53	3-	85.9	15
2-	298	4-	15.88	4-	84.7	16
3-	286	1-1970	16.40	1-1969	83.0	17
4-	264	2-	16.30	2-	82.8	18
1-1970	233	3-	15.74	3-	79.8	19
2-	224	4-	14.92	4-	73.5	20
3-	228	1-1971	14.21	1-1970	69.3	21
4-	194	2-	14.06	2-	71.5	22
1-1971	193	3-	13.76	3-	72.0	23
2-	210	4-	14.61	4-	66.9	24
3-	223	1-1972	15.06	1-1971	76.6	25
4-	238	2-	14.77	2-	80.1	26
1-1972	273	3-	15.67	3-	78.3	27
2-	287	4-	16.86	4-	79.4	28
3-	287	1-1973	17.88*	1-1972	81.8	29
4-	301	2-	18.70*	2-	86.1	30

*Forecasted from an econometric model.

Source: Process Control Company; Survey of Current Business

title headings as they appear in Figure 8.1.[1]

B(J).[2] Substituting these partial regression coefficient values into Equation 8.3, we obtain the predicting model:

$$Y_c = -188.721 + 23.6197X_2 + 1.40147X_3 - 1.77628X_4 \quad (8.4)$$

For example, to obtain \$320.346 ($x10^5$) sales in fourth quarter 1972, we substitute 18.70 for X_2 (new plant and equipment expenditures, mfg., dur.), 86.1 for X_3 (corporate profits and inventory valuation), and 30 for X_4 (time trend):

$$Y_c = -188.721 + 23.6197(18.70) + 1.40147(86.1) - 1.77628(30) = 320.346$$

If new plant and equipment expenditures had been \$1 billion higher, with the other two explanatory variables remaining the same, then predicted sales would have been 23.6×10^5 higher. Similar analyses can be made for effects of the corporate-profits-and-inventory-valuation and the time-trend variables on sales.

We note here that the relative size of the b_j coefficients is meaningful only in relation to the order of size of the input units of $X_{.j}$. That is, the b_j could be larger or smaller depending on the chosen units in which the $X_{.j}$ are expressed, because the products $b_j X_{.j}$ are summed to obtain $Y_{.c}$. The size of b_j necessary for statistical significance will be discussed in both the BETA(J) and the T-STATISTIC sections following. The positive or negative sign of the b_j coefficient, however, must be evaluated largely from the standpoint of economic causality. For example, we would expect a positive regression coefficient for almost every income explanatory variable regressed on sales and a negative regression coefficient for unemployment regressed on sales. If the sign is opposite from the expected causal relationship, then check to be sure causality justifies the observed sign and also check for nonsignificant or spurious signs due to multicollinearity.

BETA(J). In multiple regression, the partial regression coefficients, $b_2, b_3, b_4, \ldots, b_k$, are interpreted as the net influence of each causal variable on sales. Since as we indicated previously the $b_2, b_3, b_4, \ldots, b_k$ each could have different units (i.e., months, quarters, cents, dollars), it is difficult to ascertain comparative influences on sales based on the b_j's alone. Beta coefficients are one means of overcoming the scale-factor problem in the X_j's, so that direct comparisons of their relative effects on sales can be made. Beta coefficients are computed using the formula:

$$\beta_j = b_j (S_{jx}/S_y), \; j = 2, 3, 4, \ldots, k \quad (8.5)$$

where

b_j = partial (net) regression coefficient.
S_{jx} = standard deviation of explanatory variate, X_j.
S_y = standard deviation of sales.

Notice in this form, the beta coefficients are units-free numbers that are directly comparable.

Interpreting the printout, a beta coefficient is not determined (N.D.) for Y (sales). For each increase of one standard deviation in X_2 (new plant and equipment expenditures; mfg. dur.), sales, Y_c, increases by 0.906944 standard deviation units. For each increase of one standard deviation in X_4 (time trend), sales decrease by 0.458285 standard deviation units.[3] The beta coefficient for X_3 (corporate profits and inventory valuation) may be similarly interpreted. Hence, comparing all the beta coefficients, X_2 has the greatest effect on sales, followed in order by X_4 and X_3.

XBAR (J). The mean or arithmetic average for each of the four variables is calculated[4]

$$\bar{X}_j = \frac{\Sigma X_{ij}}{n} \quad (8.6)$$

where

X_{ij} = observation *i* for variable *j*.
n = number of observations.

S(J). The standard deviation for each variable is found by[5]

$$S_{jx} = \sqrt{\frac{\Sigma(X_i - \bar{X}_j)^2}{n-1}} \quad (8.7)$$

INDEX OF DETERMINATION. The index (coefficient) of determination is calculated and interpreted in a way closely resembling that used for simple regression. The value is calculated by

$$R^2_{y.234} = \frac{\text{Explained Sum of Squares}}{\text{Total Sum of Squares}} \quad (8.8)$$

Therefore, 0.774059 of the total variance in sales is explained by X_2, X_3, and X_4 in the multiple linear regression model set forth in Equation 8.4.

CORRELATION COEFFICIENT. The coefficient of multiple correlation, R, is the square root of the index of determination, R^2.

CORRELATION MATRIX. This matrix records the simple linear correlation coefficients between variables, r_{ij}. For example, Y (sales) and X_3 (corporate profits and inventory valuation) linearly covary to the extent that r = 0.729113, and so on. Notice the multiple correlation coefficient R = 0.879807 is higher than any of the simple correlation coefficients for sales with each of the explanatory variables (0.733197, 0.729113, 0.087043). Hence this comparison is one indication of how much better all of the causal variables taken collectively describe sales than each considered individually.

A second use of the *Correlation Matrix* is its role in helping identify the undesirable presence of multicollinearity (linear correlation) between explanatory variables. Although there is no currently developed statistical test of hypothesis for this purpose, L.R. Klein suggests the rule of thumb that when a simple correlation coefficient is less than the multiple correlation coefficient then any multicollinearity present between the two variables is "tolerable."[6] Using this rule as a basis, in all cases for the explanatory variables X_2, X_3, and X_4 the

Figure 8.1
Process Control Company
Multiple Regression Calculations Using
a Time-Sharing Computer

VARIABLE (J)	B(J)	BETA (J)	XBAR (J)	S (J)
1 (= Y)	-188.721	N.D.	249.933	33.548
2 (= X_2)	23.6197	.906944	15.0557	1.28817
3 (= X_3)	1.40147	.204006	78.8998	4.88343
4 (= X_4)	-1.77628	-.458285	15.5	8.65547

INDEX OF DETERMINATION (R-SQ) = .774059

CORRELATION COEFFICIENT (R) = .879807

C O R R E L A T I O N M A T R I X

1	.733197	.729113	.087043
.733197	1	.539054	.619085
.729113	.539054	1	-.079105
.087043	.619085	-.079105	1

A C T U A L V S C A L C U L A T E D

ACTUAL	CALCULATED	DIFFERENCE	PCT DIFFER
226	225.62	-.379593	-.1
245	242.2	-2.80005	-1.1
254	248.091	-5.90935	-2.3
285	261.088	-23.9124	-9.1
261	258.159	-2.84131	-1.1
249	252.5	3.49956	1.3
242	243.253	1.25325	.5
225	238.199	13.1985	5.5
235	238.302	3.30231	1.3
225	221.934	-3.06628	-1.3
216	243.393	27.3931	11.2
224	241.969	17.9693	7.4
245	267.244	22.2439	8.3
300	283.54	-16.4597	-5.8
327	295.455	-31.5447	-10.6
298	276.644	-21.3557	-7.7
286	284.768	-1.23218	-.4
264	280.822	16.822	5.9
233	261.141	28.1414	10.7
224	231.168	7.168	3.1
228	206.736	-21.2645	-10.2
194	204.5	10.4998	5.1
193	196.338	3.33827	1.7
210	207.491	-2.50887	-1.2
223	229.938	6.93797	3
238	226.217	-11.7829	-5.2
273	243.176	-29.8241	-12.2
287	271.049	-15.9514	-5.8
287	296.728	9.72803	3.2
301	320.346	19.3459	6

Figure 8.1, page 2

A N A L Y S I S O F V A R I A N C E

SOURCE	SS	DF	MS	F
REGRESSION	26135.3	3	8711.78	29.6915 @
ERROR	7628.66	26	293.41	
TOTAL	33764	29		

STANDARD DEVIATION OF ERROR TERM 17.1292

@ CONSULT F TABLE TO DETERMINE SIGNIFICANCE

VARIABLE	COEFFICIENT	STD ERROR	T STATISTIC	95% CONFIDENCE LIMITS (+,-)
2	23.6197	4.68176	5.04505	9.62569
3	1.40147	.973859	1.43909	2.00225
4	-1.77628	.588842	-3.01657	1.21066

D.F.= 26

DURBIN-WATSON STATISTIC IS 1.16476

PLOT OF RESIDUALS

HIGH RESIDUAL = 31.5447
INCREMENT FOR PLOT IS ONE PRINT SPACE = 1.15149

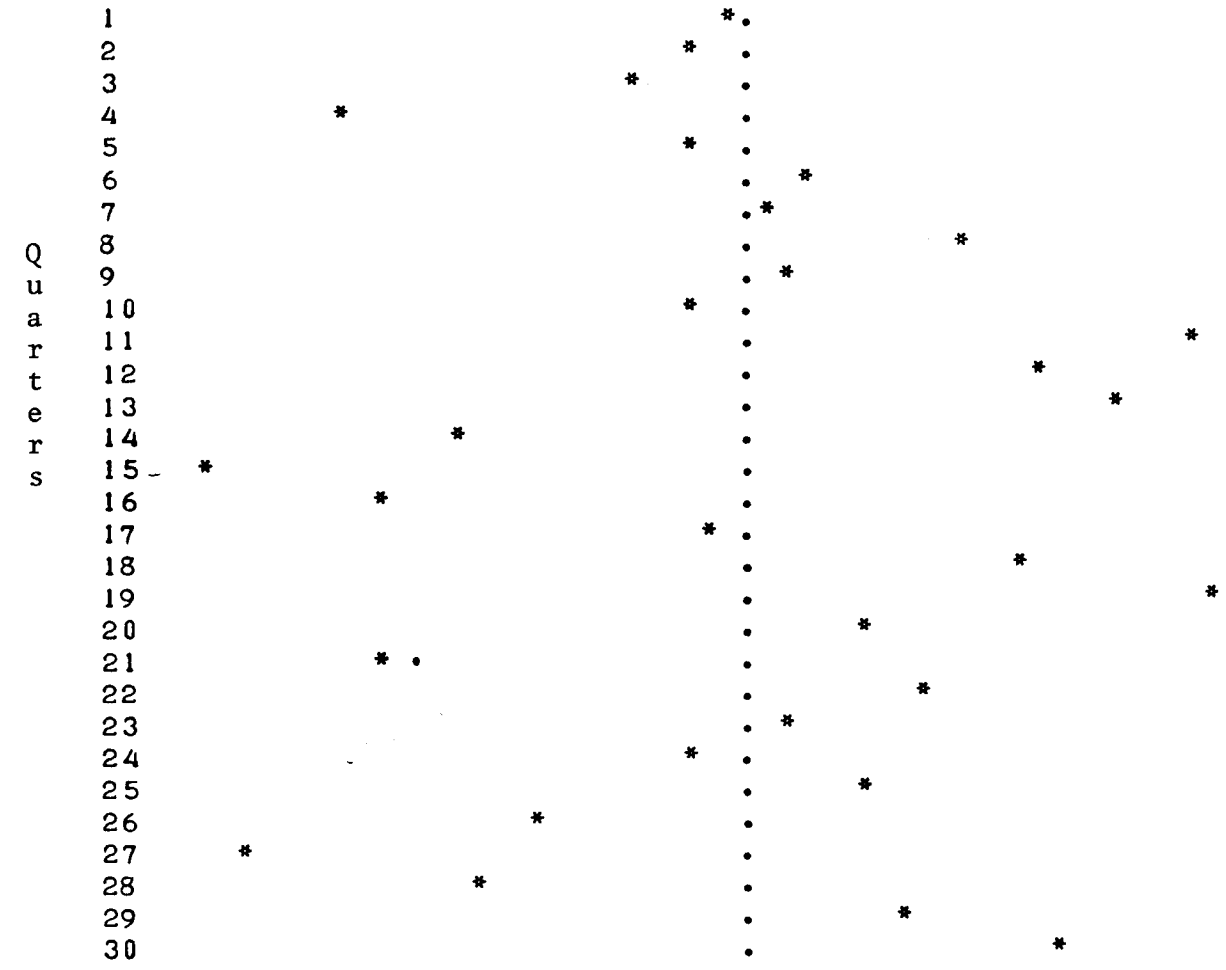


Figure 8.1, page 4

F I R S T O R D E R P A R T I A L C O R R E L A T I O N S

WITHHOLDING EFFECT OF VARIABLE 1

1	0	0	0
0	1	.0096	.8197
0	.0096	1	-.2092
0	.8197	-.2092	1

WITHHOLDING EFFECT OF VARIABLE 2

1	0	.5829	-.6871
0	1	0	0
.5829	0	1	-.6242
-.6871	0	-.6242	1

WITHHOLDING EFFECT OF VARIABLE 3

1	.5901	0	.2121
.5901	1	0	.7881
0	0	1	0
.2121	.7881	0	1

WITHHOLDING EFFECT OF VARIABLE 4

1	.8683	.7411	0
.8683	1	.7511	0
.7411	.7511	1	0
0	0	0	1

Predicted Confidence Limits

ENTER VALUE FOR X 2 ?19.695
ENTER VALUE FOR X 3 ?89.7
ENTER VALUE FOR X 4 ?31

THE 95% CONFIDENCE LIMITS ON THE EXPECTED VALUE OF Y ARE:

Y= 347.117 +- 24.2917 UPPER = 371.409 LOWER = 322.825

THE 95% CONFIDENCE LIMITS ON THE INDIVIDUAL VALUE OF Y ARE:

Y= 347.117 +- 42.7828 UPPER = 389.9 LOWER = 304.334

severity of multicollinearity is accepted as tolerable. Notice that Klein's rule is useful only for pairwise considerations. In examining more than two variables at one time, there is no definite analytical answer to the question of when we cannot accept multicollinearity as tolerable.[7]

If multicollinearity is intolerable in the sense of high simple correlations between any two explanatory variables, or in the sense of not being able to get well-determined partial regression coefficient estimates, what then are the solutions? Procedures to resolve this problem are major topics considered in Chapter 9.

ACTUAL vs. CALCULATED. This table first lists in column fashion the actual values for the dependent sales variable. Then, using Equation 8.4 and the observed values of the independent explanatory variables (from Table 8.1), the calculated Y values are presented. The differences between actual and calculated sales are recorded in the next column. Finally, the last column presents the differences as percentages of the actual sales figures.

ANALYSIS OF VARIANCE. To test the significance of R, the null hypothesis of no correlation may be accepted or rejected on the basis of the F test, where the F ratio is computed by

$$F = \frac{\text{Explained variance}}{\text{Unexplained variance}} = \frac{\text{Regression MS}}{\text{Error MS}} \qquad (8.9)$$

where

MS = SS/DF.

With a 0.05 level of significance and degrees of freedom (D.F.) equal 3 and 26, the critical F table value is 2.89. Thus, when this figure is compared to the computed F = 29.69, the hypothesis of no correlation is rejected.

STANDARD DEVIATION OF ERROR TERM. The standard deviation of error term is another name for the standard deviation of regression. In multiple regression, just as in simple regression, it is a measure of dispersion of the calculated values of the sales variable from the actual values. Thus it is a measure of average scatter of Y_i values about the line of regression and is computed by

$$S_{y.234} = \sqrt{\frac{(Y_i - Y_{ic})^2}{n - m}} = \sqrt{\frac{\text{SS Error}}{n - m}} \qquad (8.10)$$

where

Y_i = actual sales observations

Y_{ic} = calculated sales

n = number of observations

m = number of constants in the regression equation (8.4).

COEFFICIENT. This column repeats the partial (net) regression coefficients given previously under B(J).

STD ERROR. This column records the standard error for each of the b_j sampling distributions. The standard error of b_j is computed by

$$S_{jb} = \frac{S_{y.234}}{\sqrt{(X_{ij} - \bar{X}_j)(1 - R^2_{y.234})}} \quad j = 2, 3, 4 \qquad (8.11)$$

T STATISTIC. Any b_j can be tested for significance by using the *t*-test. Recall that the *t* statistic is calculated:

$$t = \frac{b_j - B_j}{S_{jb}}, \; i = 2, 3, 4, \ldots, k \qquad (8.12)$$

where

Null: $B_j = 0$, and Alternative: $B_j \neq 0$.

Hence, for X_2 the calculated *t* value using Equation 8.12 is 5.04505 and the table value at 0.05 level of significance and 26 degrees of freedom (n − m) is 2.056. Thus b_2 = 23.6197 is statistically significantly different from zero.

You may have noticed that the t-test for partial regression coefficients serves essentially the same purpose as the beta coefficients [see Beta (J)]; hence a high *t* value and the related high β_j provide similar results.

95 PERCENT CONFIDENCE LIMITS. The 95 percent confidence limits for the population partial (net) regression coefficients, (B_j), are established by

$$b_j \pm tS_{jb}, \; j = 2, 3, 4, \ldots, k \qquad (8.13)$$

For example, the confidence limits for B_2 are

23.6197 ± (2.056)(4.68176)

13.99400 to 33.24540

We conclude with 0.96 reliability that sales increase from between \$13.99400 ($x10^5$) to \$33.24540 ($x10^5$) for each billion-dollar increase in new plant and equipment expenditures (again, assuming all other explanatory variables are held constant).

DURBIN-WATSON STATISTIC. The Durbin-Watson test for autocorrelation in multiple regression is identical to that for simple regression. Checking the calculated statistic D = 1.16476 against critical table values, we conclude the test is inconclusive for positive autocorrelation at a 0.05 significance level.

PLOT OF RESIDUALS. This section of the printout is a chart of the residual errors, ϵ_i, in the same downward sequence as the time-series observations in Column 1 of Table 8.1. The dotted vertical line in the middle of Figure 8.1, page 3, is the zero residual reference line, with positive residuals plotted to the right and negative residuals to the left.

The plot of residuals is probably the most useful single tool in the entire regression printout for diagnosing inadequacies in a regression relationship. Residuals are used to analyze homoscedasticity and linearity of functional form. To illustrate both uses, it is convenient to start by analyzing simple regression and then to show how to adapt the methods to multiple regression.

Homoscedasticity (or uniform variance) related to the time sequence of observations can be visually analyzed in simple regression by determining whether the scatter of plotted residuals is approximately constant, moving from the first to the last observation.

1. If the scatter or dispersion of residuals gradually increases as observations move forward in time, then one kind of heteroscedasticity[8] is present and usually can be effectively removed by a suitable transformation of one or more variables (see Chapter 9).

2. If the scatter gradually decreases as observations move forward in time, then another kind of heteroscedasticity is present, but again this may be corrected by a different suitable transformation.

3. If the dispersion is greater in the center of the time series than at either end, then a more troublesome type of heteroscedasticity is present. This type usually means that either the explanatory variables are inadequate in the wide-dispersion part of the time series of residuals, the wrong functional regression form is being used, additional normal or dummy explanatory variables are needed, or some combination of these difficulties. (Again, see Chapter 9.)

Homoscedasticity in residuals from *multiple* regression can be analyzed analogously to the foregoing approach for simple regression. The usual starting point is to test for homogeneity in all simple regression residual patterns. Next, make necessary transformations or other changes to improve homoscedasticity, and then recalculate and plot the residuals for multiple regression. If heteroscedasticity is still present, the calculation of residuals for all possible combinations of *two* explanatory variables and the dependent sales variable is done, transforming or changing until better homoscedasticity is achieved. Eventually this process of analyzing residuals with one explanatory variable, then with all combinations of two explanatory variables, and so on, reveals the main sources of heteroscedasticity.

Linearity of functional form can be tested from the residuals in simple regression by visually determining whether the residuals form a systematic nonuniform pattern around the vertical zero reference line. For example, if the residuals form a smooth curved pattern rather than a straight line pattern, then an appropriate nonlinear regression model will fit better than a linear one.

In multiple regression, if the same smooth curved pattern of residuals appears, we also suspect nonlinearity but need to make several tests to be certain of the fact, in addition to finding where the nonlinearity lies: i.e., what combinations of variables have nonlinear relationships? This testing can be started by preparing residual plots of the dependent sales variables with each single explanatory variable, and devising appropriate nonlinear regression transformations as necessary. Then the residuals must be recalculated for the multiple regression to see if a linear path of residuals appears. If nonlinearity is still present, then the residuals of all combinations of two explanatory variables with sales must be investigated, transforming as necessary, and so forth, until a linear path of residuals for the multiple regression model is finally obtained.

Linearity can be only partly inferred by a high R or R^2, which suggests the presence of a relationship useful as a point of departure. But the only way to assure linearity in the residuals is by first examining for nonlinearity and removing it, if present, and then making reasonably certain by trial and error that no higher R or R^2 can be achieved by other conceivable types of linear and nonlinear variable transformations.

FIRST ORDER PARTIAL CORRELATIONS. Partial correlation is a measure of the degree to which that part of the variance in sales unexplained by the other explanatory variables can be explained by the introduction of an additional explanatory variable. The effect of the other explanatory variable(s) is held constant. For example, partial correlation coefficients, such as $r_{Y2.3}$, are referred to as "first order" coefficients since one independent variable is held constant: i.e., withholding the influence of X_3, what effect does X_2 have on sales? "Second order" coefficients are those where two independent variables are held constant: i.e., $r_{Y2.34}$ means withholding the influence X_3 and X_4, what effect does X_2 have on sales? In cases where no variables are held constant, the coefficients are of "zero order." Thus the order designation indicates how many independent variables are held constant in the analysis.

PREDICTED CONFIDENCE LIMITS. Most multiple linear regression programs have a predicting option. In this instance, a single-quarter sales forecast (4-1973) has been made based on forecasted inputs for the independent explanatory variables derived from a national econometric model. Sales forecasts for 4-1973 through 4-1974 are shown graphically in Figure 8.2, together with "individual" and "expected" confidence limits.

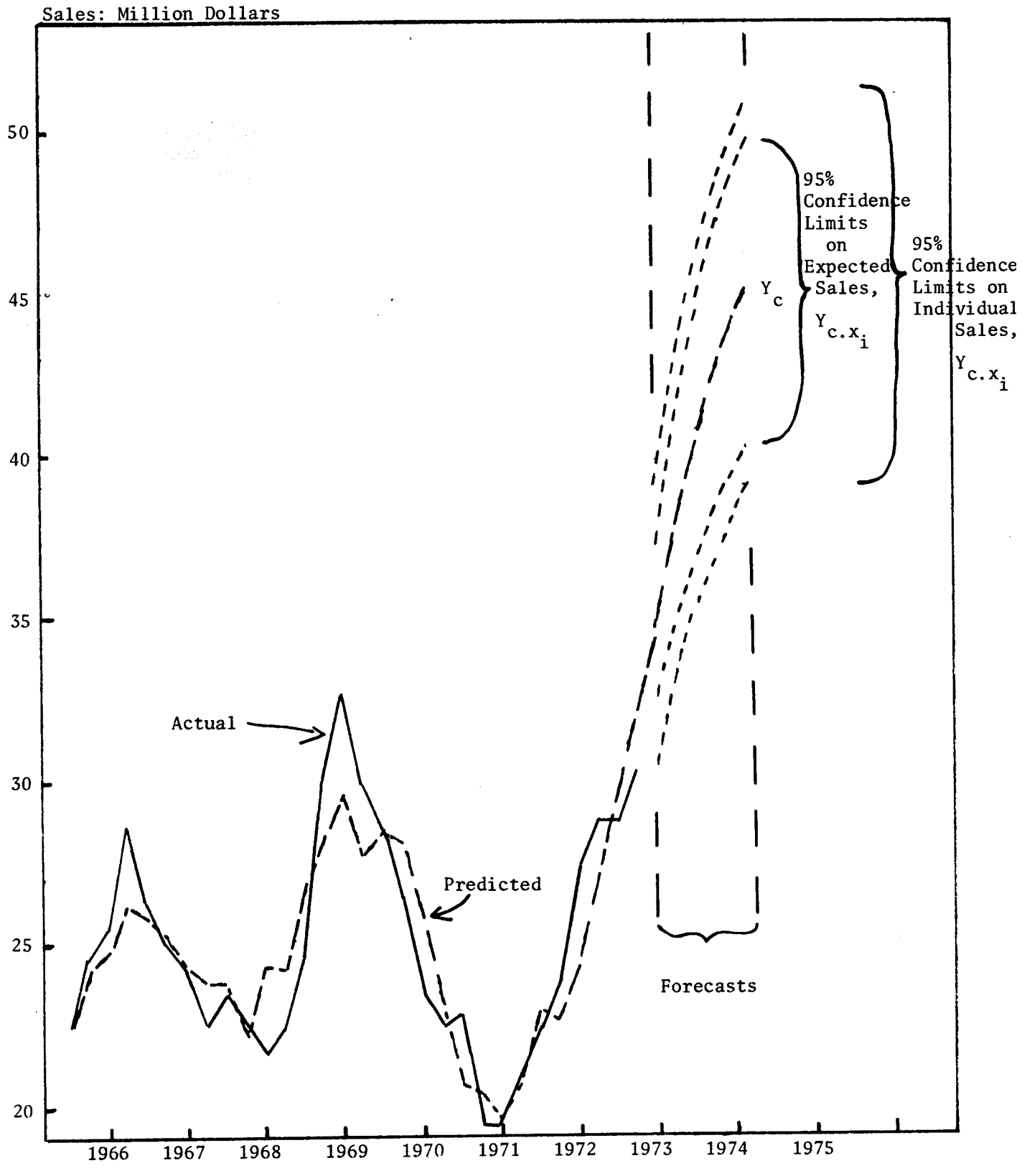
8.3 A Comparison of Standard Errors of Regression: Multiple Linear Regression vs. Time Trend

Yet another way to assess the value of multiple regression over simple regression (see discussion under *Correlation Matrix*) is by comparing standard deviations of regression. From the *Standard Deviation of Error Term* in Figure 8.1, the standard deviation of multiple regression is reported as \$17.1292 ($\times 10^5$). If the standard deviations of simple regression were computed, using each of the independent variables in turn, they would all be larger than $S_{Y.234}$. For example, sales as a function of time results in $S_{Y.4}$ = \$34.5916($\times 10^5$). We conclude that time does not do as effective a job of describing sales as does new plant and equipment expenditures, corporate profits and inventory valuation, and time considered collectively; this fact is measured by the larger standard deviation of regression in the simple regression model as compared to that for the multiple regression model.

8.4 Stepwise Multiple Regression

Before setting up a multiple regression relationship, the forecaster must give serious consideration to the issues of which explanatory variables and how many should be included. To address these two questions, multivariate linear regression analysis may be viewed as a stepwise procedure in which the sales variable is regressed successively on one explanatory variable at a time. The order in which explanatory variables are taken is determined sequentially by the most favorable partial correlation coefficients.[9] This procedure for stepwise multiple regression can best be shown by example. The computer printout in Figure 8.3 records stepwise regression results using the data developed earlier for Process Control plus one additional explanatory variable, seasonally adjusted manufacturing industrial production (X_5), considered for illustrative purposes. In this printout, sales, the dependent variable, is regressed first on the explanatory variable new plant and equipment expenditures, X_2, which has the highest zero order partial correlation coefficient or simple coefficient of correlation (see *Correlation Matrix* in

Figure 8.2
Process Control Company
Actual Sales, Predicted Sales, and Sales Forecast



Source: Table 8.1

Figure 8.3
Process Control Company
Stepwise Multiple Regression Using a Time-Sharing Computer

```
STEP  1

VARIABLE SELECTED IS ... X 2
SUM OF SQUARES REDUCED IN THIS STEP.... 18152.7
PROPORTION OF VARIANCE OF Y REDUCED.... .5377
PARTIAL F (D.F. = 1, 28 )..... 32.5667

CUMULATIVE SUM OF SQUARES REDUCED..... 18152.7
CUMULATIVE PROPORTION REDUCED..... .5377    (OF 33759.9    )

MULTIPLE CORRELATION COEFFICIENT..... .73328
F FOR ANALYSIS OF VAR. (D.F. = 1 , 28  )       32.5667
STANDARD ERROR OF ESTIMATE..... 23.6093

VARIABLE        REG. COEFF.     STD. ERR-COEF   COMPUTED T
 2               19.0941          3.3459          5.70673

INTERCEPT(A)-37.5406


STEP  2

VARIABLE SELECTED IS ... X 4
SUM OF SQUARES REDUCED IN THIS STEP.... 7367.95
PROPORTION OF VARIANCE OF Y REDUCED.... .218246
PARTIAL F (D.F. = 1, 27 )..... 24.1448


CUMULATIVE SUM OF SQUARES REDUCED..... 25520.6
CUMULATIVE PROPORTION REDUCED..... .755946  (OF 33759.9    )

MULTIPLE CORRELATION COEFFICIENT..... .869451
F FOR ANALYSIS OF VAR. (D.F. = 2 , 27  )       41.8156
STANDARD ERROR OF ESTIMATE..... 17.4687

VARIABLE        REG. COEFF.     STD. ERR-COEF   COMPUTED T
 2               28.6824          3.15222         9.09909
 4              -2.30542          .469177        -4.91374

INTERCEPT(A)-146.163
```

STEP 3

VARIABLE SELECTED IS ... X 3
SUM OF SQUARES REDUCED IN THIS STEP.... 610.392
PROPORTION OF VARIANCE OF Y REDUCED.... .180804E-01
PARTIAL F (D.F. = 1, 26)..... 2.08029

CUMULATIVE SUM OF SQUARES REDUCED..... 26131
CUMULATIVE PROPORTION REDUCED..... .774026 (OF 33759.9)

MULTIPLE CORRELATION COEFFICIENT..... .879787
F FOR ANALYSIS OF VAR. (D.F. = 3 , 26) 29.6858
STANDARD ERROR OF ESTIMATE..... 17.1294

VARIABLE	REG. COEFF.	STD. ERR-COEF	COMPUTED T
2	23.6162	4.67888	5.04741
4	-1.77597	.588563	-3.01747
3	1.40294	.972693	1.44232

INTERCEPT(A)-188.787

STEP 4

VARIABLE SELECTED IS ... X 5
SUM OF SQUARES REDUCED IN THIS STEP.... 53.0547
PROPORTION OF VARIANCE OF Y REDUCED.... .157153E-02
PARTIAL F (D.F. = 1, 25)..... .17508

CUMULATIVE SUM OF SQUARES REDUCED..... 26184.1
CUMULATIVE PROPORTION REDUCED..... .775597 (OF 33759.9)

MULTIPLE CORRELATION COEFFICIENT..... .88068
F FOR ANALYSIS OF VAR. (D.F. = 4 , 25) 21.6017
STANDARD ERROR OF ESTIMATE..... 17.4078

VARIABLE	REG. COEFF.	STD. ERR-COEF	COMPUTED T
2	22.5993	5.33998	4.2321
4	-2.06171	.907808	-2.27109
3	1.21992	1.08094	1.12858
5	.610749	1.45964	.418426

INTERCEPT(A)-219.637

Figure 8.1). This means that, when taken individually, X_2 explains the most sales variability among the four independent variables under consideration. Next, by comparing the first order partial correlation coefficients (withholding the effects of X_2), time trend, X_4, explains the largest fraction of the previously unexplained variance in sales. The process continues in this manner, comparing higher order partial correlation coefficients recalculated at each step, and X_3 and X_5 are entered in succession. Thus using all four explanatory variables results in $R_{Y.2345}$ = 0.88068. Notice that the inclusion of an additional explanatory variable will increase the R (or R^2) value if the variable has any explanatory usefulness, and since the main objective is to obtain a value of R (or R^2) as high as possible without violating the assumptions of linear regression, we may fallaciously be led to continue indiscriminately including many independent variables with the aid of computers.

As a practical matter it is not desirable to have too many variables in a regression model. Generally, only those variables that are believed to make an important contribution to the effectiveness of the predicting equation should be included, since using a large number of variables in the forecasting equation creates the inevitable problems of losing valuable degrees of freedom and of needing to obtain observations to be applied in subsequent forecasts. Moreover, interpretation of the influence of each independent variable on sales becomes quite complex.

The answer to which and how many variables to include is dictated by the practical considerations of appropriate causal analysis and the explanatory variables forecast availability. Using the stepwise approach to multiple regression is one objective way to assess the magnitude of mathematical relationship among variables. The variable manufacturing industrial production, X_5, was dropped from the analysis in Section 8.2 because of its small contribution to reducing the standard error or regression in exchange for the additional degree of freedom lost. Also, though not shown in Figure 8.3, the Durbin-Watson statistic deteriorated somewhat when X_5 was added into the analysis.

8.5 Reviewing the Assumptions of Multiple Regression Analysis

As a practical matter, primarily due to limited time and financial resources, frequently it is not possible to develop a model which satisfies all five of the assumptions of multiple regression analysis. In forecasting sales, then, what are the comparative importances and tradeoffs dictating the priorities relative to satisfying or violating underlying assumptions of regression modeling?

We suggest the following hierarchy, beginning with the most important assumption. After each title a brief review is given of how to test for the assumption's validity.

1. *Linearity*. Test for significant *R* (or R^2) using the *F* ratio; and/or test regression coefficients, the *b*, using the *t* statistic. Then test for possible improvement by the steps described in Section 8.2 under "Plot of Residuals." See Tufte for a good graphic illustration of improving the R^2 with nonlinear regression.[10]

2. *Independence*. Check for autocorrelation using the Durbin-Watson test.

3. *Homoscedasticity*. Uniform dispersion of data points is usually determined subjectively by visual inspection of a good graph. A more rigorous statistical test is accomplished by regressing the actual Y values on their respective residuals; if the resulting R (or R^2) is significant (by an F-test), then variances are not homogeneous and this assumption is violated. Graphic analyses of homoscedasticity were also described in Section 8.2 under "Plot of Residuals."

4. *Normality*. Determine graphically from the plot of residuals if the normality percentages hold for the model in question, using these bench marks:

$Y_c \pm S_{Y.X}$ includes about 68 percent
$Y_c \pm 2S_{Y.X}$ includes about 95 percent
$Y_c \pm 3S_{Y.X}$ includes about 99 percent.

5. *Non-multicollinearity*. R's (or R^2's) for combinations of simple and multiple correlations among explanatory variables using the rule of thumb that $R > 0.85$ suggest the presence of significant multicollinearity.

By far the most critical of these five assumptions are linearity and independence. Often through the process of developing a regression model that satisfies these two assumptions (see Chapter 9), the remaining three automatically become satisfied. This is not to say, however, that homoscedasticity, normality, and non-multicollinearity are unimportant, but rather their significance in the field of sales forecasting by regression is at a quantum level below that for the linearity and independence assumptions.

The following statements provided by Neter and Wasserman summarize their views on normality and non-multicollinearity in predicting by linear regression models:

"Lack of normality is not an important matter, provided the departure from normality is not of extreme form."[11]

"High multicollinearity is usually not a problem when the purpose of the regression analysis is to make inferences on the response function or predictions of new observations, provided these inferences are made within the range of observations."[12]

8.6 Advantages and Disadvantages of Causal Regression Models

The advantages of using regression models to forecast sales are several, as follows:

1. The method gives quantitative evaluation of economic and market forces, both external and internal.
2. Statistical estimates of confidence can be developed, though they are subject to many types of error.
3. Genuine causal relationships are more stable and lead to more accurate forecasts than other techniques.
4. Turning points in leading explanatory variables help predict turns in company sales.

Disadvantages of causal regression models include:

1. The method is usually time consuming and costly.
2. Historical data may be insufficient.
3. A combination of all possible errors yields wide confidence intervals.
4. A high level of statistical expertise is needed.

5. The regression equation may lack flexibility.

6. It may not be possible to forecast explanatory variables with adequate accuracy.

Footnotes

1. In computer printouts, the order depends on the way in which the program is written. Careful inspection of the output should always be made, and you should not assume that any particular order is standard.

2. B(J) here is the computer time-share terminal's equivalent form of b_j in Equation 8.3. The capital letter B is used because the terminal does not have lower case letters. Thus $B(J) = b_j$, and should not be confused with the population parameter B_j (e.g., in Equation 7.1).

3. A negative partial regression coefficient for time trend (−1.77628) does not necessarily imply that sales diminish with the passage of time. You may recall that this figure indicates "net" effect, and so in this instance serves as a slope adjustment for the multiple regression model in Equation 8.4.

4. Of course, for sales this equation more appropriately reads: $\bar{Y} = \Sigma Y_i/n$.

5. Again, for sales the correct nomenclature is:

$$S_y = \sqrt{\frac{\Sigma(Y_i - \bar{Y})^2}{n-1}}$$

6. See L.R. Klein, *An Introduction to Econometrics* (Englewood Cliffs, New Jersey, Prentice-Hall, 1962), p. 101.

7. In sales forecasting very high multicollinearity of, say, $R \geqslant 0.9$ is undesirable.

8. "Heteroscedasticity" denotes violation of the assumption of homoscedasticity (equal variances in Y_i throughout the range of $X_{.j}$ observations).

9. This "stepwise" procedure is sometimes termed "forward selection." In addition to it, Draper and Smith discuss five other algorithms for choosing variables in regression analysis; and they note the lack of a single universally accepted approach. See N.R. Draper and H. Smith, *Applied Regression Analysis* (New York, Wiley, 1966), pp. 163-177.

10. Edward R. Tufte, *Data Analysis for Politics and Policy* (Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1974), pp. 116-117.

11. John Neter and William Wasserman, *Applied Linear Statistical Models* (Homewood, Illinois, Richard D. Irwin, Inc., 1974), p. 513.

12. Ibid., p. 345.

Bibliography

Benton, William K. *Forecasting for Management.* Reading, Massachusetts: Addison-Wesley Publishing Company, 1972, ch. 3.

Bolt, Gordon J. *Market and Sales Forecasting—A Total Approach.* New York: Halsted Press Division, John Wiley and Sons, 1972, ch. 7.

Chisholm, Roger K. and Gilbert R. Whitaker, Jr. *Forecasting Methods.* Homewood, Illinois: Richard D. Irwin, Inc., 1971, ch. 7.

Chou, Ya-lun. *Statistical Analysis.* New York: Holt, Rinehart and Winston, Inc., 1969, ch. 20.

Clark, Charles T. and Lawrence L. Schkade. *Statistical Methods for Business Decisions.* Dallas, Texas: South-Western Publishing Company, 1969, ch. 18.

Draper, Norman and Harry Smith. *Applied Regression Analysis.* New York: John Wiley and Sons, Inc., 1966, ch. 4 and 6.

Huang, David S. *Regression and Econometric Methods.* New York: John Wiley and Sons, Inc., 1970, ch. 4.

Hughes, Ann and Dennis Grawoig. *Statistics: A Foundation for Analysis.* Reading, Massachusetts: Addison-Wesley Publishing Company, 1971, ch. 16.

Stockton, John R. and Charles T. Clark. *Business and Economic Statistics.* Dallas, Texas: South-Western Publishing Company, 1971, ch. 12.

Tufte, Edward R. *Data Analysis for Politics and Policy.* Englewood Cliffs, New Jersey: Prentice-Hall, 1974, p. 116.

Yamane, Taro. *Statistics: An Introductory Analysis.* New York: Harper and Row, Publishers, 1967, ch. 22.